Statistical Inference with Application to Health

Dr Mia Holley

African Institute for Mathematical Sciences
November 2025

Welcome to the course notes for **Statistical Inference with Applications to Health**. The course runs intensively from **11–29 November 2025**, with assessments designed to encourage continuous engagement with both the theoretical and applied aspects of the material. Summative assessments include quizzes, individual assignments, and a group project with a presentation.

This course is designed to provide you with a solid foundation in the principles of statistical inference, while also equipping you with the practical skills needed to implement these concepts in a modern data-science workflow. To achieve this, you will be working extensively with the `R` programming language.

Throughout the notes, you will encounter model code blocks (shown in shaded boxes). For example:

```
# Install the tidyverse package
install.packages("tidyverse")

# Load the tidyverse package
library(tidyverse)
```

You are encouraged to run each code block in your own `R` environment as you progress through the notes. Doing so will help you follow the logical flow of the analysis, understand how theoretical ideas are implemented in practice, and build confidence in using `R` for statistical inference and data analysis.

Acknowledgements

These notes are adapted from materials originally written by Mr Christopher A. Oldnall based at The University of Edinburgh.

Assessment Information

0.1 Learning Outcomes

By the end of this course, students will be able to:

1. Demonstrate a critical understanding of the fundamental concepts of statistics.

2. Apply basic statistical techniques using the statistical software `R` to solve problems in health, social and care settings.

3. Analyse and interpret health care data, using appropriate statistical tests.

4. Work both independently and collaboratively to appropriately present statistical findings on health care data.

0.2 Assessment Overview

Students must complete and submit all assessments. The final mark will be calculated from:

- Quizzes: 65%

- Individual Assignment 1: 10%

- Individual Assignment 2: 10%

- Group Project and Presentation: 15%

0.2.1 Submission Deadlines

Assessment	Weight	Date
Quiz 1	20%	Fri 14 Nov (in class)
Quiz 2	20%	Fri 21 Nov (in class)
Quiz 3	25%	Thur 27 Nov (in class)
Individual Assignment 1	10%	Sun 16 Nov, 23:59
Individual Assignment 2	10%	Sun 23 Nov, 23:59
Group Project Report	10%	Fri 28 Nov, 23:59
Group Presentations	5%	Sat 29 Nov (in class)

Contents

Chapter 1

Describing Data

1.1 What is Statistics?

The Cambridge Dictionary defines statistics in two ways:

> 1. Information based on the study of the number of times something happens or is present, or other numerical facts.
>
> 2. The science of using information discovered from studying numbers.

These definitions capture how we use statistics in everyday language. Without realising it, we rely on statistics to summarise information around us and to generalise from what we observe in order to make predictions about the future. For example, we might examine data on vaccination rates to determine how many people in a community are immunised, or whether communities with higher vaccination coverage tend to experience lower rates of infectious disease.

In formal statistical practice, questions like these fall into two broad categories:

- **Descriptive statistics** – methods used to summarise or describe observed data.
- **Inferential statistics** – methods used to make predictions or draw conclusions about a larger population based on a sample.

The need for inferential statistics arises because, in most situations, our data represent only a **sample**. From this sample we can calculate descriptive statistics such as the **mean** or **median**. However, these may not perfectly reflect the true population values. The aim of statistics is therefore twofold: to provide useful summaries of data, and to quantify the uncertainty associated with these summaries.

The mean and median are measures of central tendency. See Section 1.2.2.

It is helpful to distinguish between two types of uncertainty:

The standard deviation and inter-quartile range are measures of spread. See Section 1.2.3.

- **Uncertainty within the sample** – variation among the individuals in the observed data, often measured by the **standard deviation** or **inter-quartile range**.
- **Uncertainty from the sample** – variation in the summary statistics themselves (such as the mean or standard deviation) that would occur if we had drawn a different sample from the population.

It is this second type of uncertainty that is central to statistical inference.

1.1.1 Creating a Sample

In the context of health care, statistics is a vital tool for interpreting data and supporting decisions that affect patient care, policy, and resource allocation. Before we can analyse data, however, we must understand how it is collected. One of the most fundamental ideas in statistics is the distinction between a **population** and a **sample**.

A **population** refers to the full set of individuals or items that meet certain criteria. In health contexts, this does not necessarily mean everyone in the world. Instead, it might refer to all patients with a specific condition, all residents in a care home, or all users of a health care service.

Because it is often impractical or impossible to collect information from an entire population, researchers typically rely on a **sample**: a subset of individuals selected to represent the larger group. The key aim is for the sample to be **representative**, meaning it accurately reflects the population's characteristics. For example, to study the prevalence of diabetes in a city, it would be costly and time-consuming to survey every resident. Instead, we could select a sample of residents. If this sample is well-designed, it provides reliable estimates about the health of the wider population.

Sampling is particularly important in health care, where statistical analyses guide real-world decisions. Policy choices regarding health care funding, treatment protocols, or public health interventions often rely on sample data. For this reason, constructing well-defined, representative samples is crucial to ensure that such decisions are grounded in accurate and reliable evidence.

1.2 Distributions

After collecting data through sampling, the next step in the statistical process is to organise, summarise, and interpret it. A central concept in this process is the idea of a **distribution**.

A distribution describes how the values of a variable are spread across a range of possible outcomes. It shows whether data points cluster around certain values or are more widely dispersed. In health care, understanding distributions can reveal meaningful patterns—for example, the distribution of blood pressure readings in a population can highlight typical values and identify outliers that may signal health risks.

Distributions can be described using a range of tools, each offering a different perspective:

- **Frequency tables**: Organise data to show how often each value occurs.
- **Measures of central tendency**: The mean, median, and mode, which summarise the typical value in a dataset.
- **Measures of spread**: The range, variance, and standard deviation, which capture how dispersed the data are.
- **Visualisations**: Histograms and bar charts provide intuitive graphical summaries that make distributions easier to interpret and communicate.

In health care, distributions are particularly useful for:

1. **Defining normal and abnormal values**: For example, identifying typical ranges of blood glucose or body mass index (BMI) and flagging outliers that may require attention.
2. **Making comparisons**: Comparing distributions across groups or time periods, such as mental health scores before and after an intervention, helps evaluate effectiveness.
3. **Supporting decision-making**: Understanding the distribution of health care needs in a community guides resource allocation and service planning.
4. **Communicating findings**: Visualisations make complex distributions accessible to patients, practitioners, and policymakers, ensuring data informs practice.

In the sections that follow, we will explore these aspects of distributions in more detail, beginning with frequency tables. These provide the foundation for summarising and interpreting data, and form the first step towards deeper statistical analysis.

1.2.1 Frequency Tables

A **frequency table** lists the distinct values or categories of a variable and records how often each occurs (its frequency). By organising data in this way, it becomes much easier to detect patterns, identify common values, and spot unusual or unexpected results.
For example, consider a dataset containing the ages of patients admitted to a hospital over a month. Ages may range widely, from infants to elderly individuals. A frequency table would quickly show how many patients fall into each age group, perhaps revealing that admissions are concentrated in certain brackets.
To create a frequency table:

1. **List the values or categories**: For numerical data, group into intervals (e.g., 0–10, 11–20). For categorical data, list each category (e.g., types of health care services).

2. **Count frequencies**: Tally how many times each value or category occurs.
3. **Add relative frequencies (optional)**: Show each value's proportion of the total, often expressed as a percentage.
4. **Add cumulative frequencies (optional)**: For ordinal or numerical data, include running totals to support percentile or median calculations.

Below is an example showing the number of patients visiting a GP clinic in one week, grouped by age:

Age Group (years)	Frequency	Relative Frequency (%)	Cumulative Frequency
0–10	15	10.0	15
11–20	20	13.3	35
21–30	25	16.7	60
31–40	30	20.0	90
41–50	20	13.3	110
51–60	15	10.0	125
61–70	10	6.7	135
71+	15	10.0	150

Table 1.1: Frequency table of patients visiting a GP clinic by age group.

Here:

- The **Age Group** column shows the categories.
- The **Frequency** column counts patients in each group.
- The **Relative Frequency** column gives percentages out of 150 patients.
- The **Cumulative Frequency** column shows a running total, useful for identifying medians and quartiles.

Creating Frequency Tables in R

The steps below show how to generate a frequency table for any variable in R:

```
# Load the NHANES dataset
install.packages("NHANES")
library(NHANES)
data("NHANES")

str(NHANES)
head(NHANES)

table(NHANES$Gender)
table(NHANES$Race1, NHANES$Gender)
prop.table(table(NHANES$Race1, NHANES$Gender), margin = 1)
```

```r
# Counts per Age
age_counts <- NHANES %>%
  count(Age, name = "n")

age_counts

# Add proportions (relative frequency)
age_freq <- age_counts %>%
  mutate(prop = n / sum(n))

age_freq
```

You can also visualise the distribution with the `ggplot2` package, which is included in the package `tidyverse`:

```r
# Plot frequency of age
age_counts %>%
  ggplot(aes(x = Age, y = n)) +
  geom_col(fill = "lightblue") +
  labs(title = "Frequency␣of␣Age␣in␣NHANES",
       x = "Age␣(years)",
       y = "Count") +
  theme_minimal()
```

In the next section, we will extend frequency tables into graphical summaries, starting with histograms, which provide a powerful way to visualise distributions.

1.2.2 Measures of central tendency

Measures of central tendency are key tools for summarising data, as they identify a single value around which the data tend to cluster. In health care, they help answer questions such as: What is the typical recovery time for patients? What is the most common diagnosis in a clinic?
There are three main measures of central tendency:

- **Mean**: The arithmetic average, found by summing all values and dividing by the number of observations. Widely used but sensitive to outliers. For example, one patient with an exceptionally long hospital stay can skew the mean.
- **Median**: The middle value when data are ordered from smallest to largest. With an odd number of observations, it is the central value; with an even number, it is the average of the two central values. The median is robust to outliers and skewed distributions, making it useful for variables like income or waiting times.
- **Mode**: The most frequently occurring value. A dataset may have one mode, several modes, or none at all (if all values are unique). The mode is most helpful for categorical data, e.g. identifying the most common diagnosis or medication.

Which measure is most appropriate depends on both the type of data and its distribution:

- Use the **mean** when the data are numeric and symmetrically distributed without extreme outliers (e.g., average height).
- Use the **median** when the data are numeric but skewed, or contain outliers (e.g., income levels, length of hospital stays).
- Use the **mode** for categorical data or when the most common observation is of particular interest.

The `tidyverse` package provides straightforward tools for computing mean and median, while the mode requires a custom approach:

```
# Mean and Median of Age
NHANES %>%
  summarise(
    mean_age = mean(Age, na.rm = TRUE),
    median_age = median(Age, na.rm = TRUE))

# Mode (custom calculation)
mode_age <- NHANES %>%
  group_by(Age) %>%
  summarise(count = n()) %>%
  arrange(desc(count)) %>%
  slice(1) %>%
  pull(Age)

print(mode_age)
```

Alternatively, `summarise()` can be used.

```
# Summary of BMI
NHANES %>%
  summarise(
    median_BMI = median(BMI, na.rm = TRUE),
    mean_BMI = mean(BMI, na.rm = TRUE)
  )
```

This illustrates how each measure can offer different insights into the same variable. For symmetric distributions, mean and median will be similar, but for skewed data the median may better represent the "typical" case.

1.2.3 Measures of spread

While measures of central tendency (mean, median, mode) summarise the "centre" of a dataset, they do not reveal how widely the data values are dispersed. Two groups of patients may share the same average recovery time, but in one group recovery times might be very consistent, while in the other they vary dramatically. To capture this variation, we use measures of spread.

Measures of spread quantify the degree of variability in the data and help answer questions such as: Are patients' blood pressure readings stable, or do they fluctuate widely? How consistent are recovery times after surgery? Common measures include:

- **Range**: The difference between the maximum and minimum values. Simple to compute but very sensitive to outliers.
- **Interquartile Range (IQR)**: The range of the middle 50% of the data, calculated as Q3–Q1. This measure is robust to outliers and highlights the spread of the central portion of the distribution.
- **Variance**: The average squared deviation from the mean. Useful in theory and as a building block, but less intuitive since it is expressed in squared units.
- **Standard Deviation**: The square root of the variance, expressed in the same units as the original data. A larger standard deviation indicates greater variability; a smaller one suggests data points are clustered closely around the mean.

In health care, measures of spread are typically reported alongside measures of central tendency. For example, two hospitals may have the same mean recovery time after surgery, but a smaller standard deviation in one suggests more consistent outcomes, whereas a larger standard deviation in the other may signal outliers or differing treatment practices.

```
# Summary of BMI
NHANES %>%
  summarise(
    min_BMI = min(BMI, na.rm = TRUE),
    q1_BMI = quantile(BMI, 0.25, na.rm = TRUE),
    median_BMI = median(BMI, na.rm = TRUE),
    q3_BMI = quantile(BMI, 0.75, na.rm = TRUE),
    max_BMI = max(BMI, na.rm = TRUE),
    var_BMI = var(BMI, na.rm = TRUE),
    sd_BMI = sd(BMI, na.rm = TRUE)
  )
```

Together with the mean or median, these measures provide a fuller picture of patient data: not only the "typical" value but also how consistent or variable those values are.

1.2.4 Visualising the frequency distribution

Visualising data is one of the most effective ways to understand the distribution of values in a dataset. In health care, graphical representations help detect patterns, identify outliers, and communicate findings clearly to patients, practitioners, and policymakers.
Common methods for visualising frequency distributions include:

- **Bar Chart**: Shows frequencies of categorical or discrete data. Useful for comparing counts across categories, e.g., number of patients in different age groups.
- **Histogram**: Shows frequencies of continuous data grouped into bins. Ideal for understanding the overall shape of the distribution (symmetric, skewed, bimodal), e.g., blood pressure readings.
- **Box Plot**: Displays minimum, Q1, median, Q3, maximum, and highlights outliers. Useful for comparing distributions across groups, e.g., recovery times by hospital.
- **Density Plot**: Smoothed version of a histogram, useful for comparing continuous distributions without the discreteness of bins.

The `ggplot2` package provides flexible tools for plotting:

```
# Bar: Gender counts
NHANES %>%
  ggplot(aes(x = Gender)) +
  geom_bar(fill = "lightblue") +
  labs(title = "Participants by Gender", x = "Gender", y = "Count")

# Stacked bar chart: Race by Gender
NHANES %>%
  ggplot(aes(x = Gender, fill = Race1)) +
  geom_bar() +
  labs(title = "Participants by Gender and Race",
       x = "Gender", y = "Count", fill = "Race")

# Histogram: Age
NHANES %>%
  ggplot(aes(x = Age)) +
  geom_histogram(binwidth = 5, fill = "lightgreen", color = "black") +
  labs(title = "Histogram of Age", x = "Age (years)", y = "Count")

# Faceted Age histogram by Gender
NHANES %>%
  ggplot(aes(x = Age)) +
  geom_histogram(binwidth = 5, fill = "lightgray", color = "black") +
  facet_wrap(~ Gender) +
  labs(title = "Histogram of Age by Gender", x = "Age (years)", y = "
      Count")

# Density: BMI
NHANES %>%
  ggplot(aes(x = BMI)) +
  geom_density(fill = "lightblue", alpha = 0.6) +
  labs(title = "Density of BMI", x = "BMI", y = "Density")
```

```r
# Boxplot: BMI by Gender
NHANES %>%
  ggplot(aes(x = Gender, y = BMI)) +
  geom_boxplot(fill = "lightcoral") +
  labs(title = "BMI␣by␣Gender", x = "Gender", y = "BMI")
```

Choosing the right visualisation:

- Bar charts: Categorical comparisons.
- Histograms: Understand shape of continuous data, detect skewness or outliers.
- Box plots: Compare spread across groups and identify outliers.
- Density plots: Smooth comparison of continuous distributions.

A note on numerical accuracy

When calculating summary statistics, reporting too few significant figures can introduce rounding errors, while reporting excessively precise numbers can be misleading. To reduce cumulative rounding error:

- Perform calculations with at least two more significant figures than the final report.
- Round the final results consistently to the desired number of significant figures.
- Example: If calculating standard deviation from a variance, report variance to four significant figures and standard deviation to two.

There is no universal rule for the exact number of significant figures; consistency and transparency are key.

1.3 The normal distribution

The normal distribution is a fundamental concept in statistics, particularly in health care. It describes many natural phenomena where most observations cluster around a central value, with fewer occurring as you move away from the centre. Its characteristic bell shape (Fig. 1.1) reflects this pattern.

1.3.1 Properties of the normal distribution

Key properties of the normal distribution include:

- **Symmetry**: The left and right sides are mirror images around the mean.
- **Coinciding central measures**: Mean, median, and mode are all equal.

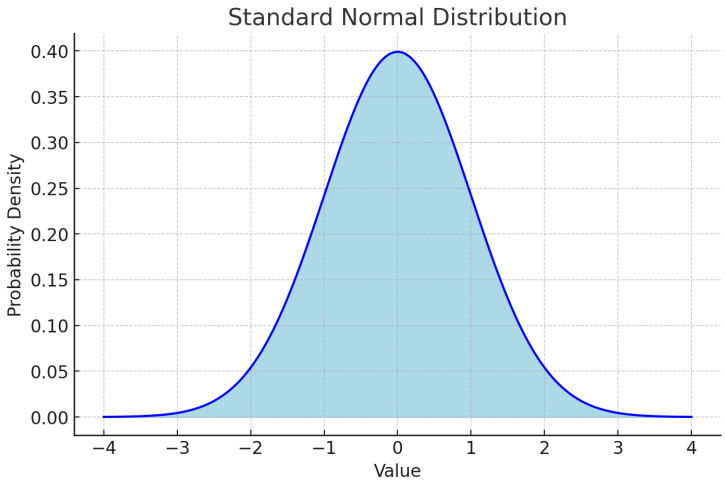

Figure 1.1: Standard normal distribution (bell curve).

- **Spread**: Standard deviation (σ) determines how concentrated or dispersed values are around the mean (μ).
- **68-95-99.7 Rule**:
  - 68% of values lie within $\mu \pm 1\sigma$
  - 95% within $\mu \pm 2\sigma$
  - 99.7% within $\mu \pm 3\sigma$

The mean locates the centre of the distribution, while the standard deviation controls the width of the bell curve. In health data, such as blood pressure or cholesterol levels, this helps identify typical values and detect abnormal results.

1.3.2 Z-scores and the standard normal distribution

A Z-score measures how many standard deviations a value is from the mean:

$$Z = \frac{x - \mu}{\sigma}$$

Where x is the value, μ the mean, and σ the standard deviation.
Z-scores allow us to:

- Standardise different datasets for comparison.
- Assess how extreme a value is relative to the population.
- Calculate probabilities for a value falling within a given range.

Example: If a population has mean cholesterol $\mu = 5.0$ mmol/L and $\sigma = 1.0$ mmol/L, a patient with $x = 6.5$ mmol/L has:

$$Z = \frac{6.5 - 5.0}{1.0} = 1.5$$

This means the patient's cholesterol is 1.5 standard deviations above the population mean.

```
# Calculate Z-scores for Cholesterol
NHANES <- NHANES %>%
  mutate(z_TotChol = (TotChol - mean(TotChol, na.rm = TRUE)) /
           sd(TotChol, na.rm = TRUE))

# Check the new column
NHANES %>% select(TotChol, z_TotChol) %>% head()
```

1.3.3 Applications of the normal distribution in health care

The normal distribution is widely used in health care because many biological measurements and health-related variables follow this pattern. Understanding and applying it helps professionals make informed decisions, identify abnormalities, and design interventions. Key applications include:

1. Clinical Measurements and Diagnostics

Many variables such as blood pressure, cholesterol, body temperature, and height are approximately normally distributed. This allows health care providers to determine "normal" ranges:

- **Blood pressure**: With a mean systolic BP of 120 mmHg and SD of 10 mmHg, 95% of individuals are expected to fall within 100–140 mmHg. Values outside this range may require further assessment.
- **Cholesterol levels**: Converting a patient's cholesterol reading to a Z-score shows how typical or extreme it is compared to the population.

2. Clinical Thresholds

Normal distributions are used to define thresholds for diagnostic tests:

- **Blood glucose**: Thresholds for diabetes or pre-diabetes often rely on population averages and standard deviations.
- **BMI**: Values are compared to population distributions to classify underweight, healthy weight, overweight, or obese.

Extreme thresholds (e.g., Z-scores > 2 or < -2) typically capture only a small percentage of the population.

3. Quality Control in health care

Statistical process control uses the normal distribution to monitor health care outcomes:

- **Patient wait times**: Comparing wait times to expected ranges highlights unusual delays.
- **Surgical success rates**: Ensures outcomes remain within acceptable ranges; deviations may indicate systemic issues.

4. Risk Assessment and Predictive Modelling
Normal distributions underpin many predictive models in health care:

- **Cardiovascular risk**: Predictive models that estimate a patient's risk of cardiovascular disease often assume a normal distribution for key variables, such as cholesterol and blood pressure. Cholesterol and blood pressure values are assessed against the normal distribution to guide preventive interventions.
- **Population health**: Models predict outcomes such as disease spread or intervention impact.

In summary, the normal distribution is widely applied in health care for clinical diagnostics, risk assessment, clinical trial analysis, and quality control. It enables health care providers to interpret patient data in a standardised way, make informed decisions, and improve the quality of care. By understanding the normal distribution, health professionals can better assess the typical range of outcomes and detect deviations that might indicate health issues or areas for improvement.

Overlaying a density curve on a histogram helps to visualise the shape and symmetry of the data:

```
ggplot(NHANES, aes(x = BMI)) +
  geom_histogram(aes(y = ..density..),
                 bins = 30,
                 fill = "lightblue",
                 color = "black") +
  geom_density(color = "red") +
  labs(title = "Distribution␣of␣BMI",
       x = "BMI",
       y = "Density")
```

1.4 Relationships between variables

Understanding how variables relate to each other is essential in health care research. By analysing these relationships, we can uncover patterns, make predictions, and understand the potential interactions that influence health outcomes. Relationships can take different forms, such as correlations, associations, or predictive models, and they can be visualised in various ways depending on the type of data involved.

Correlation measures the strength and direction of the relationship between two continuous variables. In health care, correlation analysis can help answer questions such as: Is there a relationship between age and blood pressure? or Does higher physical activity correlate with lower cholesterol levels?

One of the most commonly used measures of correlation is Pearson's correlation coefficient (r), which quantifies the linear relationship between two continuous variables. Pearson's r ranges from:

- +1: Perfect positive correlation (as one variable increases, the other increases).
- -1: Perfect negative correlation (as one variable increases, the other decreases).
- 0: No linear relationship.

The strength of the correlation is often interpreted as:

- 0.1 to 0.3: Weak correlation,
- 0.3 to 0.5: Moderate correlation,
- 0.5 to 1.0: Strong correlation.

It's important to remember that correlation does not imply causation; a high correlation between two variables doesn't mean that one causes the other. In health research, identifying whether a relationship is causal is much more complex and often requires experimental studies or clinical trials. For example, while there may be a strong correlation between higher income and better health outcomes, this doesn't necessarily mean that increasing income will directly improve health outcomes; there could be other confounding factors like access to healthcare or education.
The most effective way to visualise the correlation between two continuous variables is through a scatter plot. In a scatter plot, each point represents a pair of values for two variables, and the pattern of the points shows the nature of the relationship.

```
# Use the same filtered data for both cor.test and plot
nh_bp <- NHANES %>% drop_na(Age, BPSysAve)

# Numerical result with p-value
cor.test(nh_bp$Age, nh_bp$BPSysAve, method = "pearson")

# Visualise with regression line
nh_bp %>%
  ggplot(aes(x = Age, y = BPSysAve)) +
  geom_point(alpha = 0.5, color = "blue") +
  geom_smooth(method = "lm", se = TRUE, color = "red") +
  labs(title = "Age␣vs␣Systolic␣Blood␣Pressure",
```

```
      x = "Age␣(years)",
      y = "Systolic␣BP␣(mmHg)") +
  theme_minimal()

# Filter example: females only for AgeBP
nh_bp_female <- NHANES %>% filter(Gender == "female") %>% drop_na(Age,
    BPSysAve)
cor.test(nh_bp_female$Age, nh_bp_female$BPSysAve) # quick check

nh_bp_female %>%
  ggplot(aes(x = Age, y = BPSysAve)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = TRUE, color = "red") +
  labs(title = "Age␣vs␣Systolic␣BP␣(Females␣Only)", x = "Age␣(years)", y
      = "Systolic␣BP␣(mmHg)")
```

In health care, we often deal with different types of variables: continuous, categorical, and binary. The type of relationship being analysed influences how we visualise it. Below are some key methods for visualising relationships between different types of variables.

Continuous vs Continuous

When both variables are continuous, the scatter plot is typically used, as shown above. Another option is a heatmap if the relationship is more complex or involves a large amount of data.

Continuous vs Categorical

For relationships between a continuous variable and a categorical variable (e.g., blood pressure by gender), a box plot is a useful visualisation. A box plot shows the distribution of the continuous variable for each category, highlighting medians, quartiles, and potential outliers.

Categorical vs Categorical

For relationships between two categorical variables (e.g., smoking status vs. heart disease), a bar chart or a mosaic plot can be used to show the distribution of one categorical variable within the levels of another.

Chapter 2

Inferential Statistics

Inferential statistics allow us to make generalisations about a population based on a sample. While descriptive statistics summarise the data we have, inferential statistics enable us to infer conclusions and test hypotheses about data we haven't directly observed. This chapter introduces key concepts such as hypothesis testing, the central limit theorem, and p-values, laying the foundation for statistical tests that will be covered in future chapters.

2.1 Inferential vs Descriptive Statistics

Descriptive statistics summarise the main features of a dataset, such as central tendency (mean, median, mode), variability (range, IQR, variance, SD), and distribution shape. They describe what is observed but do not make predictions beyond the data.

```
# Summary statistics for BMI
summary(NHANES$BMI)
```

Inferential statistics allow us to draw conclusions about a population from a sample. Common goals include:

- Estimating population parameters (mean, proportion)
- Testing hypotheses to determine statistical significance

Inferential statistics are used when we want to make predictions or generalise from a sample to the larger population. Some scenarios in health care where inferential statistics are useful include:

- Determining if a new treatment is more effective than a placebo
- Estimating average recovery times after surgery
- Investigating relationships between patient age and treatment success

2.2 Hypothesis Testing

Hypothesis testing is a foundational tool in inferential statistics, allowing us to make decisions about populations based on sample data. It helps determine whether an observed effect or relationship in data is statistically significant or if it could have occurred by random chance. In health care, hypothesis testing can be used to evaluate the efficacy of treatments, compare patient outcomes, or test associations between variables.

2.2.1 Steps in Hypothesis Testing

1. **State the Hypotheses:** The null hypothesis (H_0) is a statement of no effect or no difference. The alternative hypothesis (H_1) is what we are trying to find evidence for, suggesting there is an effect or difference.
2. **Choose the Significance Level (α):** The significance level (α) is the threshold for rejecting H_0. Commonly, $\alpha = 0.05$, meaning a 5% chance of making a Type I error (rejecting H_0 when it is true).
3. **Select the Appropriate Test:** Depends on the data and research question. Common tests include t-test, Z-test, and chi-square test. In upcoming chapters, we will cover these tests.
4. **Determine the P-Value:** The p-value represents the probability of observing data as extreme as the sample, assuming H_0 is true. A small p-value ($< \alpha$) suggests rejecting H_0 in favour of H_1.
5. **Make a Decision:** If the p-value is less than the significance level ($< \alpha$), we reject H_0. If the p-value is greater than ($< \alpha$), we fail to reject H_0, meaning we do not have enough evidence to support H_1.

Example: Testing whether a licorice gargle reduces pain scores:

- H_0: $\mu_{\text{licorice}} = \mu_{\text{control}}$
- H_1: $\mu_{\text{licorice}} \neq \mu_{\text{control}}$

A t-test can quantify whether a true difference between the two means (μ) exists.

2.2.2 Types of Hypotheses

- **Two-Tailed Test:** Detects any difference in the mean between the two groups (higher or lower).
  - $H_0 : \mu_{\text{licorice}} = \mu_{\text{control}}$
  - $H_1 : \mu_{\text{licorice}} \neq \mu_{\text{control}}$
- **One-Tailed Test:** Detects difference in a specific direction.
  - $H_0 : \mu_{\text{licorice}} \geq \mu_{\text{control}}$
  - $H_1 : \mu_{\text{licorice}} < \mu_{\text{control}}$

2.2.3 Types of Errors

When conducting hypothesis tests, two types of errors may occur:

- **Type I Error (False Positive):** Rejecting H_0 when it is actually true. Probability of making a Type I error is α, the significance level. If $\alpha = 0.05$, there is a 5% chance of rejecting H_0 when it is true.
- **Type II Error (False Negative):** Failing to reject H_0 when it is false. Probability $= \beta$. The power of the test is $1-\beta$, representing the probability of correctly rejecting a false H_0.

2.3 The Central Limit Theorem, Confidence Intervals, and P-values

In inferential statistics, several key concepts underpin our ability to generalise from samples to populations. These include the Central Limit Theorem (CLT), confidence intervals, and p-values. Together, these concepts allow us to assess uncertainty in our estimates and make statistically informed decisions based on sample data.

2.3.1 Central Limit Theorem

The Central Limit Theorem (CLT) is one of the most important theorems in statistics. It states that, regardless of the population's distribution, the distribution of the **sample mean** approaches a normal distribution as the sample size increases, provided the samples are independent and drawn from the same population. This property allows us to apply normal distribution techniques even when the underlying population is not normally distributed.

Worked Example

```
# Set the random seed for reproducibility
set.seed(123)

# Replicate sampling and calculate sample means for BMI
sample_means <- replicate(
  1000,
  mean(sample(NHANES$BMI, size = 30, replace = TRUE), na.rm = TRUE)
)

# Plot the distribution of sample means
ggplot(data.frame(sample_means), aes(x = sample_means)) +
  geom_histogram(aes(y = ..density..),
                 bins = 30,
                 fill = "lightblue",
```

```
                color = "black") +
  geom_density(color = "red") +
  labs(title = "Distribution␣of␣Sample␣Means␣of␣BMI␣(Central␣Limit␣
      Theorem)",
       x = "Sample␣Mean␣BMI",
       y = "Density") +
  theme_minimal()
```

2.3.2 Confidence Intervals

A confidence interval (CI) is a range of values likely to contain the true population parameter with a specified level of confidence. The most common is the 95% CI, meaning we are 95% confident the interval contains the true population parameter.

For a population mean (μ), the CI is calculated as:

$$\text{CI} = \bar{x} \pm z \cdot \frac{\sigma}{\sqrt{n}}$$

where

- $\bar{x}$ is the sample mean,
- z is the Z-value associated with the desired confidence level (e.g., 1.96 for 95%),
- σ is the sample standard deviation,
- n is the sample size.

A 95% confidence interval means that if we were to take 100 different samples and compute a confidence interval for each sample, we would expect the true population mean to lie within the interval in approximately 95 of those samples. It does not mean that there is a 95% chance that the population mean is in any one particular interval.

Worked Example

```
# Calculate mean and standard error of BMI
mean_BMI <- mean(NHANES$BMI, na.rm = TRUE)
stderr_BMI <- sd(NHANES$BMI, na.rm = TRUE) / sqrt(nrow(NHANES))

# Calculate 95% confidence interval
conf_interval <- mean_BMI + c(-1.96, 1.96) * stderr_BMI
print(conf_interval)
```

In this example, the confidence interval provides a range within which we can be 95% confident that the true mean pain score of the population lies.

2.3.3 P-values

A p-value is the probability of obtaining results as extreme as (or more extreme than) those observed, assuming the null hypothesis (H_0) is true. It measures how compatible the data are with H_0.
In practice:

- **Small p-value ($< \alpha$):** Evidence against H_0; we reject H_0 in favour of H_1.
- **Large p-value ($\geq \alpha$):** Insufficient evidence against H_0; we fail to reject it.

Note: A small p-value does not prove H_0 is false, and a large p-value does not prove it is true — it only reflects the strength of evidence.

2.4 Assumptions of Parametric Tests

Parametric tests, such as t-tests and ANOVA, rely on key assumptions about the data. Violating these assumptions can lead to misleading results.

2.4.1 Normality

The normality assumption refers to the sampling distribution of the test statistic, not necessarily the raw data. For small samples, check the data distribution. For larger samples, the CLT ensures the sample mean is approximately normal. Methods to check normality:

- **Histogram/Density Plot:** Visual check of data distribution.
- **Q-Q Plot:** This plot shows how the sample data compare to a normal distribution. Points near the line suggest approximate normality.
- **Shapiro-Wilk Test:** This test assesses whether the sample data are normally distributed. p-value > 0.05 suggests data are not significantly different from normal.

For smaller sample sizes, where the normality of the sampling distribution cannot be guaranteed by the Central Limit Theorem, you may consider using transformations (e.g., log transformation) to make the data more normal. Alternatively, you can use non-parametric tests, such as the Mann-Whitney U test, which do not assume normality. We will not look into these methods in this module.

2.4.2 Homogeneity of Variance

Homogeneity of variance (also known as homoscedasticity) assumes that the variance within each group being compared is approximately equal. This assumption is important for tests like the t-test and ANOVA, where we compare means across groups. If the variances across groups are not equal, the test may

overestimate or underestimate the significance of the results, leading to erroneous conclusions. When homogeneity of variance is violated, alternative tests such as the Welch's t-test can be used, as they are more robust to unequal variances.

We can use visualisations (like box plots) and statistical tests (Levene's test) to check for homogeneity of variance. In the case of box plots, if the boxes (inter-quartile ranges) in the box plot are roughly the same height, the variances are likely to be equal. We'll look more carefully at Levene's test in Chapter 3.

2.4.3 Independence

The assumption of independence requires that the observations in the data are independent of each other. This means that one data point should not influence another. Independence is crucial for both t-tests and ANOVA.

If observations are not independent (e.g., repeated measures or paired data), traditional parametric tests like the t-test may not be appropriate. Instead, we would use tests like the paired t-test or mixed models, which account for the dependence between observations. While there is no single universal test for independence in all contexts, there are statistical methods to test for dependence or 'auto-correlation' in specific situations, especially in time series data, repeated measures, or hierarchical data. One test we will look at in Chapter 3 is the χ^2 test. Ensuring proper data collection procedures and avoiding repeated measures without accounting for them is essential.

2.5 The inferential statistics pipeline.

The inferential statistics pipeline provides a structured framework for analysing data and drawing conclusions about a population based on a sample. This approach ensures consistency and reliability in statistical analyses. The key steps are as follows:

1. Define the research question and hypotheses. Clearly articulate the research question and state the null hypothesis (H_0) and alternative hypothesis (H_1). The null hypothesis typically represents no effect or no difference, while the alternative hypothesis represents the effect or difference of interest (see Section 2.2).

2. Choose an appropriate statistical test. Select the statistical test that aligns with the type of data and the research question. For instance:

- Use a t-test to compare means (Section 3.1.2).
- Use a chi-square test to examine relationships between categorical variables (Section 3.3).

- Use ANOVA to compare means across multiple groups (see Chapter 4).

3. Perform the test and calculate test statistics. Carry out the statistical test to obtain the test statistic (e.g., t-value, F-value) and p-value. The test statistic quantifies the evidence against H_0, while the p-value indicates whether the observed result is statistically significant (see Section 2.3).

4. Interpret the results. Compare the p-value to the significance level (α, typically 0.05):

- If $p < \alpha$: Reject H_0. There is sufficient evidence to support H_1.
- If $p \geq \alpha$: Fail to reject H_0. There is insufficient evidence to support H_1.

Additionally, confidence intervals can provide further insight into the precision of the estimate (Section 2.3).

5. Check assumptions of the chosen test. Ensure that the data meet the assumptions required for the test, such as normality, independence, and homogeneity of variances (Section 2.4). If assumptions are violated, consider transformations or alternative non-parametric tests.

6. Draw conclusions and report findings. Summarise the findings in the context of the research question. Discuss whether the results are practically significant and highlight any limitations of the analysis. This pipeline ensures a systematic approach to inferential statistics, linking the theoretical foundations of hypothesis testing with practical decision-making in health care.

Chapter 3

Statistical Testing

Statistical testing is a fundamental aspect of inferential statistics, allowing researchers to draw conclusions about populations based on sample data. In health care, where we often deal with incomplete or limited data, the ability to infer broader trends or differences from smaller sample sizes is crucial for decision-making, policy development, and clinical practices.

The core of statistical testing lies in evaluating hypotheses: *educated guesses* about population parameters or relationships between variables. These hypotheses are tested using statistical methods that provide a framework for determining whether observed data are consistent with the hypothesis or deviate significantly enough to suggest an alternative explanation. Statistical tests quantify uncertainty, assess risk, and support evidence-based conclusions.

Statistical testing also ensures rigour and reproducibility. In health care, where the stakes are high—such as evaluating new treatments—accurately determining whether observed differences are due to chance or a true effect can influence critical decisions. Proper statistical testing mitigates biases, reduces errors, and establishes confidence in research outcomes.

In this chapter, we explore several types of statistical tests commonly used in health care research, focusing on tests for equality of means, equality of variances, and variable dependence. Each area addresses specific research questions when comparing groups, assessing variability, or exploring relationships between variables.

In health care, the application of statistical testing is essential for a variety of reasons, some of which we list below.

- **Clinical trials**: Statistical tests are fundamental in assessing the effectiveness of new treatments, medications, or interventions. By comparing the outcomes of treatment groups with control groups, statistical testing helps

determine whether an intervention produces significant improvements in patient outcomes.

- **Public health decision-making**: Statistical testing informs decisions about public health policies and resource allocation. For instance, when deciding whether to implement a new health policy, statistical evidence is needed to ensure that the policy would likely result in positive health outcomes for the target population.
- **Health inequality research**: In studies of health inequalities, statistical tests are used to assess whether disparities in health outcomes across different demographic groups (e.g., based on socioeconomic status, race, or geographic location) are statistically significant or likely due to random variation.
- **Diagnostic accuracy**: Statistical testing also aids in determining the sensitivity and specificity of diagnostic tools or tests, ensuring that medical professionals can make reliable and accurate diagnoses based on empirical data.

The three primary areas covered in this chapter—equality of means, equality of variances, and variable dependence—each serve different research purposes. Tests for equality of means help us compare average outcomes across groups, such as treatment versus control groups in a clinical trial. Tests for equality of variances are used when it's important to ensure that the variability in outcomes is similar between groups. Finally, tests for variable dependence assess whether two categorical variables, for example smoking status and heart disease occurrence, are statistically related.

Understanding when and how to apply these statistical tests is crucial for health care professionals. Throughout this chapter, we will delve into these concepts with practical examples, providing a clear understanding of the mechanics of each test, its assumptions, and its applications in real-world scenarios.

3.1 Testing for Equality of Means

When comparing average outcomes between two or more groups, tests for equality of means determine whether observed differences are statistically significant or could have occurred by chance. Common tests include the *Z-test* and *T-test*, chosen based on sample size and whether population variance is known.

3.1.1 Z-Test

The Z-test compares the means of two independent groups when the population variance is known and the sample size is large ($n > 30$). The sampling distribution of the mean is assumed normal under the Central Limit Theorem.

The test statistic is:

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}}$$

where

- $\bar{X}_1, \bar{X}_2$ are the sample means,
- σ^2 is the known population variance,
- n_1, n_2 are the sample sizes.

The null hypothesis (H_0) assumes $\mu_1 = \mu_2$, and the alternative (H_1) assumes $\mu_1 \neq \mu_2$.

Application: Comparing mean recovery times between patients receiving two treatments, assuming a sufficiently large sample size and known population variance. Reject H_0 if the calculated Z exceeds the critical value for the chosen significance level (e.g., 1.96 for a 95% confidence level).

3.1.2 T-Test

When the population variance is unknown, the t-test is used. Variants include the one-sample, independent samples, and paired samples t-tests.

One-Sample T-Test

Tests whether the mean of a single sample differs from a known or hypothesised population mean:

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

where $\bar{X}$ is the sample mean, μ the hypothesised mean, s the sample standard deviation, and n the sample size.

Application: Assessing whether the average height in a sample differs from a national average (e.g., 162 cm).

Independent Samples T-Test

Compares the means of two independent groups. Welch's t-test, shown below, does not assume equal population variances and is the recommended approach in most practical applications:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where $\bar{X}_i$ and s_i^2 are the sample mean and variance, and n_i the sample size.

Application: Comparing mean blood pressure between patients receiving medication and those in a control group, to determine if observed differences are statistically significant.

Paired Samples T-Test

Used when the same subjects are measured under two conditions or at two time points. The test evaluates whether the mean difference between paired observations differs significantly from zero:

$$t = \frac{\bar{d}}{s_d/\sqrt{n}}$$

where $\bar{d}$ is the mean of differences, s_d their standard deviation, and n the number of pairs.

Application: Assessing changes in mobility before and after a rehabilitation programme, or in crossover trials where the same participants receive both treatment and control.

Decision Rules

Once the test statistic is calculated, there are two common ways to decide whether to reject the null hypothesis H_0:

- **Critical value method:** Compare $|t_{\text{calc}}|$ with the critical value from the t-distribution table at the chosen significance level α and degrees of freedom (df). Reject H_0 if $|t_{\text{calc}}| > t_{\alpha,df}$.
- **p-value method:** Compute the p-value corresponding to the test statistic. Reject H_0 if $p \leq \alpha$.

Worked Example (One-Sample T-Test)

Suppose a sample of $n = 25$ students has a mean height of $\bar{X} = 165$ cm with standard deviation $s = 10$. Test at $\alpha = 0.05$ whether the population mean differs from $\mu = 162$.

$$t = \frac{165 - 162}{10/\sqrt{25}} = \frac{3}{2} = 1.50$$

Degrees of freedom: $df = 25 - 1 = 24$.

Critical value method: From the t-table, $t_{0.05,24} \approx 2.064$. Since $1.50 < 2.064$, we fail to reject H_0.

p-value method: The two-tailed p-value for $t = 1.50, df = 24$ is about 0.147. Since $0.147 > 0.05$, we again fail to reject H_0.

Conclusion: At the 5% significance level, there is insufficient evidence to conclude that the mean height differs from 162 cm.

Worked Example (Independent Samples T-Test)

A study compares mean systolic blood pressure between two independent groups: a treatment group ($n_1 = 20$) receiving a new medication, and a control group ($n_2 = 22$). The sample means and standard deviations are:

$$\bar{X}_1 = 118.4, \quad s_1 = 9.5, \qquad \bar{X}_2 = 124.1, \quad s_2 = 8.7$$

Test at $\alpha = 0.05$ whether there is a significant difference in mean blood pressure between the two groups.

$$t = \frac{118.4 - 124.1}{\sqrt{\frac{9.5^2}{20} + \frac{8.7^2}{22}}} = \frac{-5.7}{\sqrt{4.512 + 3.442}} = \frac{-5.7}{2.84} = -2.01$$

Degrees of freedom (Welch's approximation):

$$df = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}} \approx 39.$$

Critical value method: From the t-table, $t_{0.05,39} \approx 2.02$. Since $|t_{\text{calc}}| = 2.01 < 2.02$, we fail to reject H_0.

p-value method: The two-tailed p-value is approximately 0.051. Since $p > 0.05$, we again fail to reject H_0.

Conclusion: At the 5% significance level, there is insufficient evidence to conclude that the two treatments produce different mean blood pressures.

Worked Example (Paired Samples T-Test)

Researchers measure pain intensity in 12 patients before and after a licorice-gargle treatment. The differences (Before − After) in pain scores are:

$$d_i = \{1.2, 0.8, 2.0, 1.5, 1.1, 0.5, 1.3, 1.8, 0.9, 1.6, 1.0, 0.7\}$$

The mean and standard deviation of the differences are:

$$\bar{d} = 1.20, \qquad s_d = 0.45.$$

Test at $\alpha = 0.05$ whether the treatment significantly reduces pain.

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} = \frac{1.20}{0.45/\sqrt{12}} = \frac{1.20}{0.130} = 9.23$$

Degrees of freedom: $df = 12 - 1 = 11$.

Critical value method: From the t-table, $t_{0.05,11} = 2.201$. Since $9.23 > 2.201$, we reject H_0.

p-value method: The two-tailed p-value is less than 0.001.

Conclusion: There is strong evidence that the licorice-gargle treatment reduces mean pain scores.

3.1.3 Assumptions of the Z-Test and T-Test

Both the Z-test and T-test have specific assumptions that must be met for the results to be valid:

- Data approximately normally distributed, especially for small samples.
- Samples are independent (for independent T-tests and Z-tests).
- For Z-test: population variance known, sample size large.
- For T-test: sample variances approximately equal (use Welch's T-test if not).

Violating these assumptions may require non-parametric alternatives like the Mann-Whitney U test (independent samples) or Wilcoxon signed-rank test (paired samples), which are beyond the scope of this course.

3.2 Testing for Equality of Variances

When testing for equality of variances, we aim to determine whether the variability within different groups is the same. This is crucial because many statistical tests, such as analysis of variance (ANOVA) and t-tests, assume that the groups being compared have similar variances. Unequal variances can affect the accuracy of these tests, potentially leading to incorrect conclusions. One of the most widely used methods for testing equality of variances is **Levene's test**.

3.2.1 Levene's Test

Levene's test is commonly used to assess the equality of variances across multiple groups. It tests the null hypothesis that the variances in different groups are equal, making it an essential diagnostic step before conducting parametric tests like the t-test or ANOVA. Levene's test is particularly useful because it is robust against deviations from normality, meaning it works well even if the data are not normally distributed.

Levene's test calculates the absolute deviations of each data point from its group mean and then performs an ANOVA on these absolute deviations. By focusing on deviations from the group means, Levene's test reduces the impact of outliers, making it more reliable than tests that assume normality.

Levene's test produces a p-value. If the p-value is **less than 0.05** ($p < 0.05$), this indicates that the variances are significantly different, and the assumption of equal variances is violated. If the p-value is **greater than or equal to 0.05** ($p \geq 0.05$), the test suggests that the variances are equal, and the assumption holds.

3.2.2 Why Levene's Test is Important

Levene's test helps ensure that the assumptions of parametric tests are met. If the variances are found to be unequal, using standard versions of the t-test or ANOVA can lead to inaccurate conclusions. In such cases, alternative methods like Welch's t-test should be considered, as it does not assume equal variances.

You can easily run Levene's test in R using the `car` package. Below is an example of how to perform the test:

```
# Install the car package if not already installed
install.packages("car")

# Load the car package
library(car)

# Example of running Levene's Test for equality of variances
leveneTest(BMI ~ Gender, data = NHANES)
```

The test will return a p-value, which you can use to determine whether the variances are equal.

Levene's test is a key tool for ensuring the validity of parametric tests like ANOVA and t-tests. By checking the equality of variances, we can ensure that our statistical tests are reliable and that we draw accurate conclusions from the data. Always run Levene's test before assuming equal variances in your analysis.

3.3 Testing for Variable Dependence (χ^2 Test)

The χ^2 (Chi-square) test is one of the most widely used statistical tests for determining whether there is a significant association between two categorical variables. In health care, this test is frequently used to investigate relationships between different categories, such as the association between treatment type and patient recovery outcomes, or between demographic factors and health conditions.

The χ^2 test evaluates whether the observed frequencies in each category differ significantly from the frequencies that would be expected if the variables were independent. In other words, the test checks whether one variable depends on another or whether the two variables are independent of each other. There are two main types of χ^2 tests:

1. **The χ^2 Test for Independence:** This test is used when you want to examine whether there is an association between two categorical variables in a contingency table.
2. **The χ^2 Goodness-of-Fit Test:** This test is used to determine whether the observed frequency distribution of a categorical variable matches an expected distribution.

In this course, we focus on the χ^2 test for independence, as it is commonly used to assess relationships between two categorical variables.

3.3.1 How the χ^2 Test for Independence Works

The χ^2 test for independence compares the observed frequencies in each category to the expected frequencies. The expected frequencies are calculated under the assumption that there is no association between the variables, i.e., they are independent.

Contingency Table: The data are typically summarised in a contingency table, which displays the frequencies of occurrence for each combination of the categories of the two variables. For example, suppose we want to test whether there is an association between smoking status (smoker/non-smoker) and the incidence of lung disease (yes/no). The contingency table would look something like Table 3.3.1.

	Lung Disease	**No Lung Disease**	**Total**
Smoker	40	20	60
Non-Smoker	10	30	40
Total	50	50	100

Table 3.1: Contingency table of smoking status and lung disease.

The χ^2 test evaluates whether the distribution of observed frequencies (e.g., 40 smokers with lung disease) is significantly different from the expected frequencies, which are calculated assuming no relationship between smoking status and lung disease.

Expected Frequencies: The expected frequency for each cell in the contingency table is calculated using the formula:

$$E = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}$$

For example, for the cell representing smokers with lung disease:

$$E = \frac{60 \times 50}{100} = 30$$

This value represents the frequency we would expect if smoking status and lung disease were independent.

Chi-Square Statistic: Once the expected frequencies are calculated, the χ^2 statistic is computed as follows:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Where:

- O is the observed frequency in each cell,
- E is the expected frequency in each cell.

The χ^2 statistic aggregates the squared differences between observed and expected frequencies, weighted by the expected frequencies. A larger χ^2 value indicates a greater discrepancy between the observed and expected values, suggesting a stronger association between the variables.

The degrees of freedom (df) for a χ^2 test in a contingency table are calculated as:

$$df = (r - 1) \times (c - 1)$$

Where r is the number of rows and c is the number of columns in the contingency table. The degrees of freedom are necessary to determine the critical value of the χ^2 statistic from the χ^2 distribution table.

3.3.2 Interpreting the Results

After calculating the χ^2 statistic and determining the degrees of freedom, the next step is to compare the χ^2 statistic to the critical value from the χ^2 distribution. This can be done by using a χ^2 distribution table or by obtaining a p-value.

- **If the p-value is less than the significance level** (typically 0.05), we reject the null hypothesis and conclude that there is a significant association between the two variables.
- **If the p-value is greater than or equal to the significance level**, we fail to reject the null hypothesis and conclude that there is no significant association between the variables.

Assumptions of the χ^2 Test:

- The data must be in the form of counts or frequencies.
- The categories must be mutually exclusive, meaning each observation falls into one and only one category.
- The expected frequency for each cell should be at least 5 for the test to be valid. If this assumption is violated, other methods, such as combining groups, should be considered.

You can run a χ^2 test for independence in R using the `chisq.test()` function. Here's an example of how to perform the test with a contingency table:

```
# Create a contingency table of Smoking Status and Physical Activity
contingency_table <- table(NHANES$SmokeNow, NHANES$PhysActive)

# Run the Chi-square test of independence
chisq.test(contingency_table)
```

In this example, the function will return the χ^2 statistic, the degrees of freedom, and the p-value.

The χ^2 test for independence is a powerful tool for examining the association between two categorical variables. By comparing the observed frequencies to the expected frequencies under the assumption of independence, we can determine whether there is a significant relationship between the variables. It is widely applicable in health care for analysing relationships between treatments, outcomes, demographics, and more.

Chapter 4

Analysis of Variance

In many statistical applications, we need to compare the means of several groups to determine whether they differ significantly. For example, in health care research, we might compare treatment effects, intervention outcomes, or demographic differences. While a two-sample t-test works for comparing two groups, it becomes inefficient and error-prone when comparing more than two simultaneously.

Analysis of Variance (ANOVA) addresses this issue by comparing multiple group means within a single framework. It partitions the total variation in the data into components reflecting *between-group* and *within-group* variability, allowing us to test whether observed differences are statistically significant or likely due to random variation.

This chapter introduces the logic of ANOVA, beginning with the decomposition of variation (sum of squares), the calculation of mean squares, inference using the F-ratio, post-hoc testing, and the assumptions underlying valid results.

4.1 Why ANOVA is Necessary

Performing multiple t-tests across several groups has major drawbacks:

- **Increased Type I error:** Each t-test carries a risk of a false positive; multiple tests compound this risk.
- **Reduced efficiency:** ANOVA provides a single, comprehensive test instead of many pairwise comparisons.
- **Loss of overall insight:** ANOVA assesses whether any group differences exist and, if significant, post-hoc tests can identify which groups differ.

In essence, ANOVA tests for differences among several group means simultaneously, offering both statistical rigor and efficiency.

4.2 Partitioning Variation: The Sum of Squares

A central goal in statistics is to understand how data vary. ANOVA achieves this by partitioning total variation into two components: *between-group* and *within-group* variation.

For example, suppose we compare exam scores of three different classes. Differences in average scores reflect between-group variation, while differences among students within each class reflect within-group variation. ANOVA determines whether the observed between-group variation is large enough to conclude that the group means differ beyond random chance.

4.2.1 Components of Variation

Variation is quantified using the *sum of squares* —the sum of squared deviations of data points from a reference point. ANOVA distinguishes three related sums of squares:

- **Total Sum of Squares (SST):** Overall variation of all observations from the grand mean.
- **Between-Groups Sum of Squares (SSB):** Variation due to differences between group means.
- **Within-Groups Sum of Squares (SSW):** Variation within each group, around its own mean.

Formally:

1. **Total Sum of Squares (SST):**

$$SST = \sum_{i=1}^{N} (X_i - \overline{X})^2$$

   where X_i is the ith observation, $\overline{X}$ is the grand mean, and N is the total sample size.

2. **Between-Groups Sum of Squares (SSB):**

$$SSB = \sum_{j=1}^{k} n_j (\overline{X}_j - \overline{X})^2$$

   where k is the number of groups, n_j the size of group j, and $\overline{X}_j$ its mean.

3. **Within-Groups Sum of Squares (SSW):**

$$SSW = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (X_{ij} - \overline{X}_j)^2$$

   where X_{ij} is the ith observation in group j.

$$\boxed{SST = SSB + SSW}$$

This decomposition forms the foundation of ANOVA, allowing us to separate systematic group differences from random variation.

4.3 Mean Squares and Degrees of Freedom

The *mean square* (MS) is the average variation associated with each source of variability. It is obtained by dividing each sum of squares by its respective degrees of freedom (df).

- **Mean Square Between Groups (MSB):**

$$MSB = \frac{SSB}{df_B} = \frac{SSB}{k-1}$$

  Represents the average variation between group means.

- **Mean Square Within Groups (MSW):**

$$MSW = \frac{SSW}{df_W} = \frac{SSW}{N-k}$$

  Represents the average variation within groups.

Large differences between group means yield a high MSB, while high variability within groups leads to a large MSW. Comparing these quantities forms the basis of the ANOVA F-test.

4.3.1 Degrees of Freedom

Degrees of freedom adjust for the number of independent pieces of information used:

- Between groups: $df_B = k - 1$
- Within groups: $df_W = N - k$

These are essential for computing unbiased variance estimates and for referencing the F-distribution in hypothesis testing.

4.4 Inference in ANOVA

4.4.1 The F-Ratio

To test whether group means differ significantly, we compute the *F-ratio*:

$$F = \frac{MSB}{MSW}$$

A large F value indicates that the between-group variation is substantially greater than the within-group variation, suggesting that not all group means are equal.

4.4.2 Interpreting the F-Ratio

The calculated F statistic is compared to a critical value from the F-distribution with (df_B, df_W) degrees of freedom.

- If F exceeds the critical value (or if $p < 0.05$), reject H_0: at least one group mean differs.
- If F is smaller (or $p \geq 0.05$), fail to reject H_0: observed differences may be due to chance.

4.4.3 Hypotheses

- **Null hypothesis (H_0):** All population means are equal.

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

- **Alternative hypothesis (H_1):** At least one group mean differs.

Worked Example

```
# Filter NHANES data for adults
dataset1 <- NHANES %>%
  filter(Age > 18) %>%
  select(SleepHrsNight, Work) %>%
  drop_na(SleepHrsNight, Work)

# Descriptive statistics of sleep hours by work category
dataset1 %>%
  group_by(Work) %>%
  summarise(
    mean_sleep = mean(SleepHrsNight, na.rm = TRUE),
    sd_sleep = sd(SleepHrsNight, na.rm = TRUE),
    var_sleep = var(SleepHrsNight, na.rm = TRUE)
  )

# Box plot of sleep duration by work status
ggplot(dataset1, aes(x = Work, y = SleepHrsNight, fill = Work)) +
  geom_boxplot() +
  labs(title = "Hours␣of␣Sleep␣by␣Work␣Status",
       x = "Work␣Status",
       y = "Hours␣of␣Sleep␣per␣Night") +
  theme_minimal()
```

```
# One-way ANOVA: effect of work status on sleep hours
ANOVA1 <- aov(SleepHrsNight ~ Work, data = dataset1)
summary(ANOVA1)
```

4.5 Post-hoc Tests

A significant ANOVA result indicates that not all means are equal, but it does not reveal which groups differ. To identify specific differences, we perform *post-hoc tests*, which make pairwise comparisons while controlling Type I error.
Common post-hoc methods include:

- Tukey's HSD
- Bonferroni correction
- Scheffé's method

We will not cover these methods in this course.

4.6 Assumptions of ANOVA

Like most inferential tests, ANOVA relies on several assumptions.

4.6.1 Independence

Observations must be independent. Ensure random sampling and avoid repeated measurements unless using a repeated-measures ANOVA.

4.6.2 Homogeneity of Variance

Variances should be similar across groups. Use Levene's test to check:

```
# Levenes test for equal variances
leveneTest(SleepHrsNight ~ Work, data = dataset1)
```

4.6.3 Normality of Residuals

Residuals should follow an approximately normal distribution. Check via histograms, Q-Q plots, or the Shapiro–Wilk test.

```
# Visual checks for normality
dataset1 %>%
  ggplot(aes(x = SleepHrsNight)) +
  geom_histogram(bins = 10, color = "white", fill = "lightblue") +
  facet_wrap(~Work, ncol = 1, scales = "free_y")
```

```r
ggplot(dataset1, aes(sample = SleepHrsNight)) +
  stat_qq() +
  stat_qq_line()

# ShapiroWilk test (sampled subset for large dataset)
sample <- slice_sample(dataset1, n = 4000)
shapiro.test(sample$SleepHrsNight)
```

4.6.4 Handling Violations

- **Independence:** Must not be violated.
- **Unequal variances:** Use Welch's ANOVA.
- **Non-normality:** For large samples, ANOVA is robust; otherwise, use non-parametric alternatives (e.g., Kruskal–Wallis test).

Summary

- ANOVA compares multiple group means using the F-statistic.
- Total variation is decomposed as $SST = SSB + SSW$.
- Mean squares standardize variation: $MSB = SSB/(k-1)$ and $MSW = SSW/(N-k)$.
- A large $F = MSB/MSW$ indicates significant group differences.
- Post-hoc tests identify which means differ, provided assumptions are satisfied.

Chapter 5

Correlation and Regression

In statistics, we are often interested in quantifying and modelling relationships between variables. For example, in health care, we might want to understand whether physical activity is related to blood pressure, or whether age affects recovery time after surgery. To study such relationships between continuous variables, we use two key tools: *correlation* and *regression*.

Correlation measures the strength and direction of the relationship between two variables. It tells us whether variables tend to move together (positive correlation) or in opposite directions (negative correlation). However, correlation only indicates association, not causation.

Regression, on the other hand, models the relationship between a dependent variable (outcome) and one or more independent variables (predictors). In its simplest form, univariate linear regression represents this relationship as a straight line, enabling prediction of one variable from another.

Together, correlation and regression form the foundation of applied statistics, allowing us to quantify relationships and make predictions. This chapter introduces both concepts with practical examples in R.

5.1 Correlation

5.1.1 What is Correlation?

As introduced earlier, correlation measures how two continuous variables change in relation to one another. Specifically, it tells us:

- **Strength of the relationship:** How strongly the variables are related.
- **Direction of the relationship:** Whether the variables increase or decrease together (positive correlation), or if one increases while the other

decreases (negative correlation).

A positive correlation means that as one variable increases, the other tends to increase as well; a negative correlation means that as one increases, the other tends to decrease.

5.1.2 Correlation Coefficients

The correlation coefficient, denoted by r, ranges between -1 and 1:

- $r = 1$: Perfect positive correlation — both variables increase together exactly.
- $r = -1$: Perfect negative correlation — one variable increases while the other decreases exactly.
- $r = 0$: No linear correlation — no consistent linear pattern between the variables.

The most widely used method for measuring the linear relationship between two continuous variables is **Pearson's correlation coefficient**. It assumes that both variables are measured on a continuous scale and that the relationship between them is linear. The formula is:

$$r = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum(X_i - \overline{X})^2 \sum(Y_i - \overline{Y})^2}}$$

where X_i and Y_i are the data points for variables X and Y, and $\overline{X}$ and $\overline{Y}$ are their respective means.

When using Pearson's r, the following assumptions should hold:

- The relationship between the variables is linear.
- Both variables are approximately normally distributed.
- The data are free from influential outliers.

Even if r is close to zero, a non-linear relationship may still exist. Always examine a scatterplot before interpreting r.

Spearman's correlation coefficient is a non-parametric alternative that measures the strength and direction of a monotonic (not necessarily linear) relationship between two variables. It is suitable for ordinal or non-normally distributed data. We will not focus on it in this module.

Worked Example

```r
# Correlation between diastolic and systolic blood pressure (adult males)
subsetNHANES <- NHANES %>%
  filter(Gender == "male", Age >= 18) %>%
  drop_na(BPDiaAve, BPSysAve) %>%
  select(BPDiaAve, BPSysAve)

# Pearson correlation test
subsetNHANES %>%
  summarise(
    Pearson_r = cor.test(BPDiaAve, BPSysAve, method = "pearson")$estimate
        ,
    p_value = cor.test(BPDiaAve, BPSysAve, method = "pearson")$p.value
  )

# Visualise the correlation
ggplot(subsetNHANES, aes(x = BPDiaAve, y = BPSysAve)) +
  geom_point(alpha = 0.4, color = "steelblue") +
  geom_smooth(method = "lm", se = FALSE, color = "red", linewidth = 1) +
  labs(
    title = "Diastolic␣vs␣Systolic␣Blood␣Pressure␣(Adult␣Males)",
    x = "Diastolic␣Blood␣Pressure␣(mmHg)",
    y = "Systolic␣Blood␣Pressure␣(mmHg)"
  ) +
  theme_bw()
```

This code filters the NHANES dataset for adult males and computes Pearson's correlation between diastolic and systolic blood pressure, returning both the correlation coefficient and its p-value. The scatterplot provides a visual check of the positive linear association.

5.1.3 Interpreting Correlation Coefficients

Interpret r in terms of both its direction and strength:

- **Positive correlation ($r > 0$):** As one variable increases, the other also increases.
- **Negative correlation ($r < 0$):** As one variable increases, the other decreases.
- **No correlation ($r = 0$):** No linear relationship between the variables.

Typical guidelines for interpreting the magnitude of Pearson's r are:

- $|r| = 0.1 - 0.3$: Weak correlation
- $|r| = 0.3 - 0.5$: Moderate correlation
- $|r| > 0.5$: Strong correlation

Always remember that **correlation does not imply causation**. A strong correlation does not mean that one variable causes the other to change—it only indicates association.

5.1.4 Hypothesis Testing for Correlation

We can formally test whether the observed correlation is statistically significant:

- **Null Hypothesis (H_0):** $r = 0$ (no correlation)
- **Alternative Hypothesis (H_1):** $r \neq 0$ (non-zero correlation)

The test statistic is:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

which follows a t-distribution with $n-2$ degrees of freedom. If the p-value is below the chosen significance level (e.g., $\alpha = 0.05$), we reject H_0 and conclude that a statistically significant correlation exists.

5.2 Univariate Linear Regression

Having established how correlation measures association, we now move to regression, which enables prediction.

5.2.1 Concept and Model

Univariate linear regression models the relationship between one independent variable (X) and one dependent variable (Y) with a straight line:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

where:

- Y — dependent (outcome) variable,
- X — independent (predictor) variable,
- β_0 — intercept (predicted Y when $X = 0$),
- β_1 — slope (expected change in Y per unit change in X),
- ϵ — error term (difference between observed and predicted Y).

5.2.2 Ordinary Least Squares (OLS)

OLS estimates the regression line by minimising the sum of squared residuals:

$$\text{Residual} = Y_i - \hat{Y}_i, \qquad \text{Minimise} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

This ensures the fitted line is as close as possible to the observed data points.

5.2.3 Interpreting Coefficients

- **Intercept (β_0):** Predicted value of Y when $X = 0$.
- **Slope (β_1):** Expected change in Y for a one-unit increase in X.

A positive slope indicates a positive relationship; a negative slope indicates an inverse relationship.

Worked Example

```
# Fit a linear regression model
lmodel <- lm(BPSysAve ~ BPDiaAve, data = subsetNHANES)

# Model summary (coefficients, R-squared, F-statistic)
summary(lmodel)

# 95% confidence intervals for coefficients
confint(lmodel)
```

This model predicts systolic blood pressure based on diastolic blood pressure. The output provides estimates for the intercept and slope, their significance (via p-values), and model fit statistics.

5.2.4 Goodness of Fit: The Coefficient of Determination (R^2)

$$R^2 = 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \overline{Y})^2}$$

R^2 indicates the proportion of variance in Y explained by X. Values close to 1 suggest a strong fit, while values near 0 indicate a weak fit.

5.2.5 Residuals and Diagnostics

Residuals should be randomly scattered around zero with no visible pattern. A structured pattern implies that the model may not capture the relationship appropriately.

5.3 Inference for Normal Linear Regression

After fitting a regression model, we use inference to test whether predictors have statistically significant relationships with the outcome.

Each coefficient has an associated standard error, quantifying the uncertainty of the estimate. To test whether a coefficient differs significantly from zero, we use a t-test:

- **t-value:** Number of standard errors the coefficient is away from zero.
- **p-value:** Probability of observing such a coefficient under $H_0 : \beta = 0$.

If the p-value is less than 0.05, we conclude that the predictor has a statistically significant effect on Y.

5.4 Assumptions of Normal Linear Regression

The validity of regression results depends on key assumptions:

5.4.1 Linearity

The relationship between X and Y should be linear. A scatterplot can help verify this. If curvature is visible, consider transforming variables or using polynomial regression.

5.4.2 Constant Variance (Homoscedasticity)

Residuals should have constant variance across all fitted values. A funnel shape in a residual plot suggests heteroscedasticity.

```
# Residuals vs Fitted values plot
ggplot(data.frame(fitted = fitted(lmodel), resid = resid(lmodel)),
       aes(x = fitted, y = resid)) +
  geom_point() +
  geom_hline(yintercept = 0, color = "red") +
  labs(x = "Fitted␣values",
       y = "Residuals",
       title = "Residuals␣vs␣Fitted␣Values") +
  theme_bw()
```

5.4.3 Independence of Observations

Each observation should be independent of others. If data include repeated measurements or clusters, use models like mixed-effects regression.

5.4.4 Normality of Residuals

Residuals should be approximately normally distributed.

```
# Check normality of residuals
res <- resid(lmodel)

# Histogram of residuals
ggplot(as_tibble(res), aes(x = value)) +
  geom_histogram(bins = 30, fill = "lightblue", color = "black") +
  labs(title = "Histogram␣of␣Residuals",
```

```
      x = "Residuals",
      y = "Count") +
  theme_bw()

# QQ plot of residuals
ggplot(as_tibble(res), aes(sample = value)) +
  stat_qq() +
  stat_qq_line(color = "red") +
  labs(title = "Normal QQ Plot of Residuals") +
  theme_bw()

# ShapiroWilk test for normality
shapiro.test(res)
```

5.4.5 Outliers

Outliers are data points that fall significantly outside the trend formed by the majority of the data. In linear regression, outliers can have a disproportionately large influence on the estimated coefficients, potentially skewing the slope and intercept and thus leading to biased or misleading results. Outliers may arise from data entry errors, measurement inconsistencies, or natural variability. However, regardless of their source, it is essential to identify and consider their impact on the model.

In some cases, excluding outliers is appropriate, especially if they represent errors. In other cases, robust regression techniques may be used to reduce their impact without exclusion. Visual inspection, particularly through scatterplots or residual plots, is an effective way to identify outliers. Additionally, statistical measures, such as Cook's distance, can help assess the influence of individual points. Addressing outliers is important to ensure the regression model accurately reflects the underlying relationship without undue influence from extreme values.

5.5 Conclusion

Correlation and regression together provide a framework for understanding and predicting relationships between variables. Correlation quantifies association, while regression allows prediction and inference. However, regression results are only reliable when model assumptions are satisfied and when influential points are properly addressed. Checking these diagnostics ensures that your conclusions are both accurate and meaningful.

Course Conclusion

Throughout this course, we have covered a range of foundational statistical concepts, techniques, and practical tools essential for analysing data in health care. We began with the basics of describing and summarising data, explored different types of statistical tests, and delved into methods for examining relationships between variables. Along the way, we've seen how understanding data distributions, testing hypotheses, and building regression models can provide meaningful insights into complex datasets, helping us make informed, evidence-based decisions in a health context.

As you continue your journey with statistics, remember that data analysis is both a science and a skill. The concepts introduced in this course lay a strong foundation, but expertise in statistics and data science grows with practice, critical thinking, and real-world application. Each dataset is unique, and developing a robust analytical approach often involves a cycle of exploration, testing, and refinement. For example, when applying regression analysis, the next steps may involve experimenting with multiple predictors, exploring non-linear relationships, or handling larger, more complex datasets.

Looking ahead, here are a few suggestions for building on what you've learned:

- Explore More Advanced Statistical Techniques: As your comfort with foundational methods grows, consider learning more advanced techniques such as logistic regression, multivariable analysis, time series analysis, or machine learning methods. These tools can expand your ability to answer increasingly complex questions in health care.

- Apply and Practise in Real-World Settings: Whenever possible, seek opportunities to work with actual health data. Applying your skills to real datasets—whether in research projects, internships, or work placements—will deepen your understanding and build confidence in interpreting and communicating your findings.

- Develop Your Coding Skills in R or Other Statistical Software: Familiarity with R has been a central part of this course. Enhancing your coding skills will enable you to automate repetitive tasks, streamline analyses, and work

efficiently with larger datasets. You might also consider exploring other statistical software or programming languages, such as Python, that are commonly used in data science.

- Strengthen Your Understanding of Research Methodology: Understanding the principles of study design, sampling, and ethical considerations in data collection will support your statistical skills and enable you to critically evaluate research in health care.
- Engage with the Statistical Community: Participating in forums, attending workshops or conferences, and joining professional networks can keep you updated with the latest statistical approaches and connect you with others in the field.

This course is just the beginning of a lifelong learning journey in statistics. Whether you pursue further study or apply these skills directly in your career, statistics will remain an invaluable tool for understanding the world, informing decisions, and contributing to evidence-based improvements in health care.

Recommended Books

Douglas Altman. *Practical Statistics for Medical Research*. Chapman and Hall/CRC, 1990. ISBN 9780412276309.

Martin Bland. *An Introduction to Medical Statistics*. Oxford University Press, 4 edition, 2015. ISBN 9780199589920.

Theodore Colton. *Statistics in Medicine*. Little, Brown and Company, 1974. ISBN 0316153203.

Betty Kirkwood and Jonathan Sterne. *Essential Medical Statistics*. Blackwell Science, 2 edition, 2003. ISBN 9780865428713.